Chemical reaction completion

A hybrid rule-based and language model-based approach

by

Matthijs van Wijngaarden

to obtain the degree of Master of Science
at Delft University of Technology,
to be defended publicly on
November 13th, 2023 at 14:00.

Student number: 4271785
Project duration: March 2023 – November 2023

Thesis committee:

Prof. dr. ir. M.J.T. Reinders	TU Delft, Responsible advisor
Dr. J.M. Weber	TU Delft, Daily supervisor
G. Vogel MSc.	TU Delft, Daily co-supervisor
Dr. M. Khosla	TU Delft

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Chemical reaction completion: a hybrid rule-based and language model-based approach

Matthijs van Wijngaarden

October 30, 2023

Abstract

Large chemical reaction databases often suffer from incompleteness, such as missing molecules or stoichiometric information. Concurrently, numerous computational models are being developed in predictive chemistry that rely on reaction databases and would hugely benefit from complete reaction equations. Also, research in sustainable chemistry often focuses on automated mass balance tasks, which require a full reaction to properly evaluate. In this work, we present a hybrid approach for computational completion of reaction equations. Specifically, we combine a rule-based method and a machine learning (ML) model to complete reactions. The rule-based approach constructs a balance of atoms and charge on either side of the reaction in an attempt to find missing molecules. We tailor the pre-trained transformer model on the chemical language domain to take partial reactions as inputs and predict missing molecules. Furthermore, we present a novel approach to measure the correctness of our model, which is useful when we apply it to the uncurated dataset and the ground-truth is unknown.

Introduction

Chemical reactions can be described as the transformation of one set of of molecules (reactants) to another set of molecules (products). The network of organic chemistry is a graphical representation of this system, where nodes (the molecules) are connected to each other by edges (their relevant chemical reactions) [1]. This network is fundamental for research and development in cheminformatics. As new molecules are synthesized and new reactions are discovered, the size of this network grows. Harnessing this continuous growth of chemical information requires a proper store of data. Traditionally, reactions were manually recorded, which is a tedious process and a slow solution to a growing problem. In recent years, efforts have been made to automate this process using data mining techniques on chemical patents [2]. As a result, large amounts of chemical reaction data have been made freely accessible, and have been widely used by researchers and chemists for a variety of tasks [3, 4].

A notable limitation to current chemical reaction datasets is the high rate of incompleteness among reactions stored in databases. A chemical reaction is considered complete when all reacting components of the reaction are included with correct stoichiometry. In other words, a complete reaction contains the correct amount of every molecule which directly contributes to the atom flow of the reaction. This definition of completeness excludes chemical context molecules such as solvents, catalysts or other types of non-atom-contributing reagents. Completeness of a reaction in this definition can be deduced by counting the atoms as well as the charge of the reacting and product molecules, respectively. If the same amount of every involved atom type is observed for reactants and products, and the cumulative charge of reactant molecules equals the cumulative charge of product molecules, the reaction is considered complete. Figure 1 shows two examples of unaltered, incomplete reactions. The first reaction is a typical example of a reaction missing a small byproduct, in this case hydrogen chloride (HCl). The second reaction is an example of erroneous stoichiometry, as synthesizing the product molecule requires two bicyclic aromatic compounds. An analysis on various subsets of the orig-

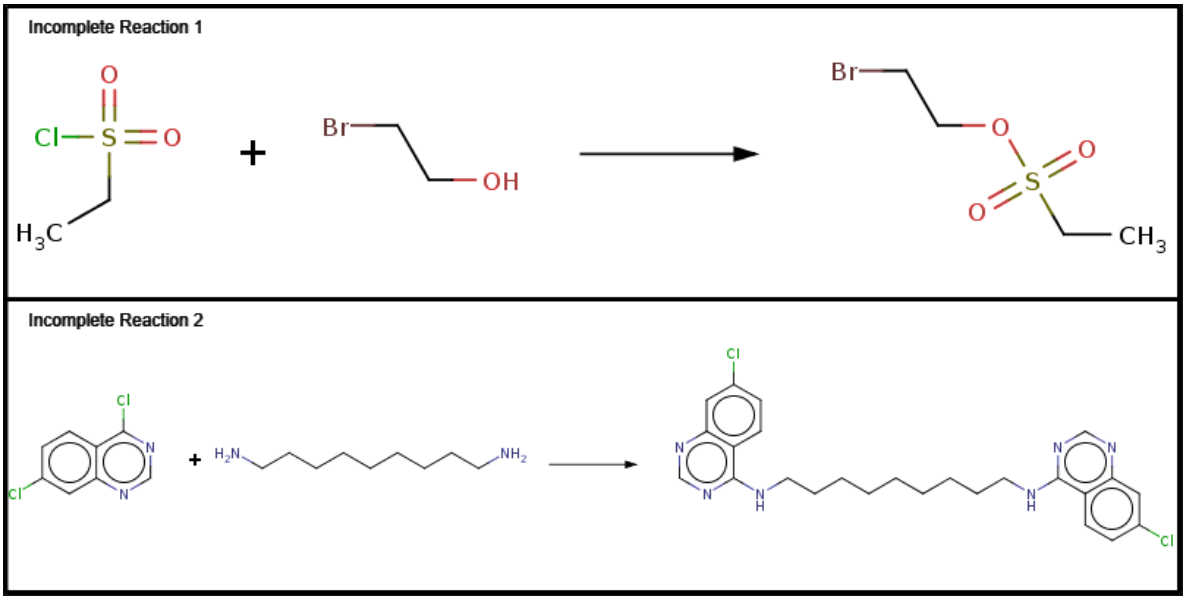

Figure 1: Two typical examples of incomplete reactions found in chemical reaction databases. **Reaction 1** shows the synthesis of a compound by fusing two reactant molecules. The chlorine atom of the sulfuric compound reacts with the hydroxyl (-OH) group of the other reactant, releasing both the chlorine atom of the first reactant, and the hydrogen atom of the second reactant. Both the H atom and Cl atom are missing and form a hydrogen chloride (HCl) molecule, which is not pictured. **Reaction 2** shows the synthesis of a compound by reacting the bicyclic aromatic compound to the long carbon-chain molecule. It is obvious that two molecules of the first reactant are needed to synthesize the product molecule, despite only one molecule being recorded in the data.

inal, patent-mined dataset known as the USPTO[1] (United States Patent and Trade Office) dataset showed that as few as 2%-3% of reactions are naturally balanced. Consequentially, 97%-98% of reactions are imbalanced.

An earlier paper on reaction incompleteness theorized two possible reasons for this common observation [5]. Firstly, when chemists discover new reactions, the synthesized product is usually the main goal, while smaller side molecules such as co-reactants and byproducts are omitted. The second reason is that incompleteness can be the result of a noisy, imperfect data-mining technique. Patents might describe certain reacting molecules in other parts of the document. Additionally, the variation in structure of patents makes it difficult to correctly identify all contributing molecules [5].

Incomplete reactions pose a problem across various fields of cheminformatics. Predictive chemistry is a rapidly developing field of research which includes, but is not limited to, forward reaction prediction, retrosynthesis, reaction yield prediction and reaction condition prediction [6, 7, 8, 9]. Most modern methods developed in these fields are data-driven and directly rooted in the aforementioned, highly incomplete chemical datasets. While state-of-the-art models have produced impressive results, they are based on incomplete data. Curated chemical datasets which contain complete reactions for such models could produce more reliable, correct results.

The problem of having incomplete reaction data extends to the field of 'Green Chemistry'. This area of study focuses on the development of sustainable solutions in chemical engineering [10]. A recent review regarding sustainable chemistry addresses certain issues faced and categorizes them in three broad categories: data, assessment metrics, and decision-making [3]. The authors identify that, in particular, the data category causes the most bottlenecks, where data *completion* is mentioned specifically. This is due to most sustainable chemistry research focusing on finding an optimized reaction pathway based on a multi-objective framework. Optimizing sustainability of a chemical process looks at multiple factors, such as the use of biological feedstock molecules, waste streams and emission numbers. However, in order to remain competitive in the industry, economic factors must also be taken into account when designing chemical reaction pathways [11]. Chemical process engineers employ automated mass bal-

[1] https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873

ance tasks using chemical reaction data to discover an optimal reaction pathway. Early-stage evaluation of such pathways use an evaluation function based on a multi-objective framework [3]. Unfortunately, many of these objectives are difficult to measure due to incompleteness in the available reaction data. For example, missing byproduct molecules directly contribute to an erroneous waste stream analysis, while missing co-reactants potentially voids proposed reactions due to being high-cost or environmentally unfriendly. Guaranteeing completeness of reaction data would propel the efficiency of automated reaction pathway optimization.

Developments are being made to address the reaction incompleteness problem. Reaction data accessibility has historically been attributed to private institutions or companies, such as Reaxys™[2] (henceforth referred to as Reaxys) and CASREACT[3]. Recently, an effort has been made to aggregate chemical reaction data in an open source database. The Open Reaction Database (ORD) is a promising open-source initiative, supported by multiple institutions [12]. This database prioritizes a standardized data structure, including reaction completeness. Its goal of providing a consistent data representation for downstream use of predictive chemistry tasks is emphasized. However, since only a fraction of the organic chemistry data space is recorded in this manner, automated reaction completion needs to be considered.

Automated reaction completion methods attempt to balance incomplete chemical reactions using code. Methods in this field of research can be divided into two categories: rule-based and model-based. Rule-based methods attempt to balance an incomplete reaction by encoding rules into an algorithm. The specific rules may vary in approach. One such approach is by use of a condensed graph of a reaction (CGR). The CGR is a superposition of the graph representation of the reactants and products in a potentially incomplete reaction. By analyzing bond changes and deducing missing atoms, a CGR-based study partially curated reaction databases [13]. A different rule-based technique is to formulate the problem as an atom-balancing problem [5, 14]. Any missing atoms on either side of the reaction can then be attributed, through simple logical rules and limited chemical deduction, to a known small molecule regarded as 'helper species'.

The second category of automated reaction completion is based on machine learning. Graph model-based and language model-based methods (LM-based methods) have been explored in various tasks in predictive chemistry [7, 15]. A transformer-based LM, the Molecular Transformer, has shown accuracy above 90% in the forward synthesis prediction problem [6]. The transformer is a state-of-the-art LM that uses an encoder-decoder architecture with multi-head attention layers and positional feed forward layers [16]. The self-attention mechanism allows the model to capture long-range dependencies in text-based data. Language models have traditionally been designed for natural language tasks, such as neural machine translation or text interpretation. However, their strength can be used in other problem domains as well, such as chemistry. Reactions are represented in text format, like sentences, where molecules can be considered 'words', and atoms its 'letters'. Using the Molecular Transformer as a base, a study taught the LM to predict missing molecules in incomplete reactions, rather than predicting product molecules [17]. This version of an autoregressive generative transformer model is purely sequence-to-sequence, wherein it inputs a chemical reaction in string format, and outputs what it believes is the set of missing molecules in string format. Another approach of model-based reaction completion is ChemBalancer, which uses the BERT architecture: a modified transformer without a decoder [5]. ChemBalancer was inspired by BERT's success at discovering the missing word in a natural language sentence. They shifted this problem domain towards the language of chemistry, where missing words in a sentence are akin to missing molecules in a reaction. Notably, this paper introduces the hybrid rule-based and model-based approach to curating reaction datasets [5].

In this paper, we propose a unique hybrid method to the reaction completion problem. We present a method that is partially based on encoded, algorithmic rules, and partially based on a trained transformer-based LM. By first applying a rule-based method on an incomplete reaction dataset, we curate the dataset partially, generating a subset of reactions which are complete. This generated subset is used as ground-truth to train our LM. By feeding the model partialized reactions, it learns the context of the reaction completion problem, and once fully trained, can be applied to the uncurated subset.

In this paper, we explore and answer a number of pressing questions regarding the reaction incompleteness problem. We identify that a majority of reaction data can be readily curated using only rule-based algorithm. However, reactions that are

[2]https://supportcontent.elsevier.com/RightNow%20Next%20Gen/Reaxys/New_RX_FactSheet_Jul_2018.pdf

[3]https://www.cas.org/support/documentation/reactions

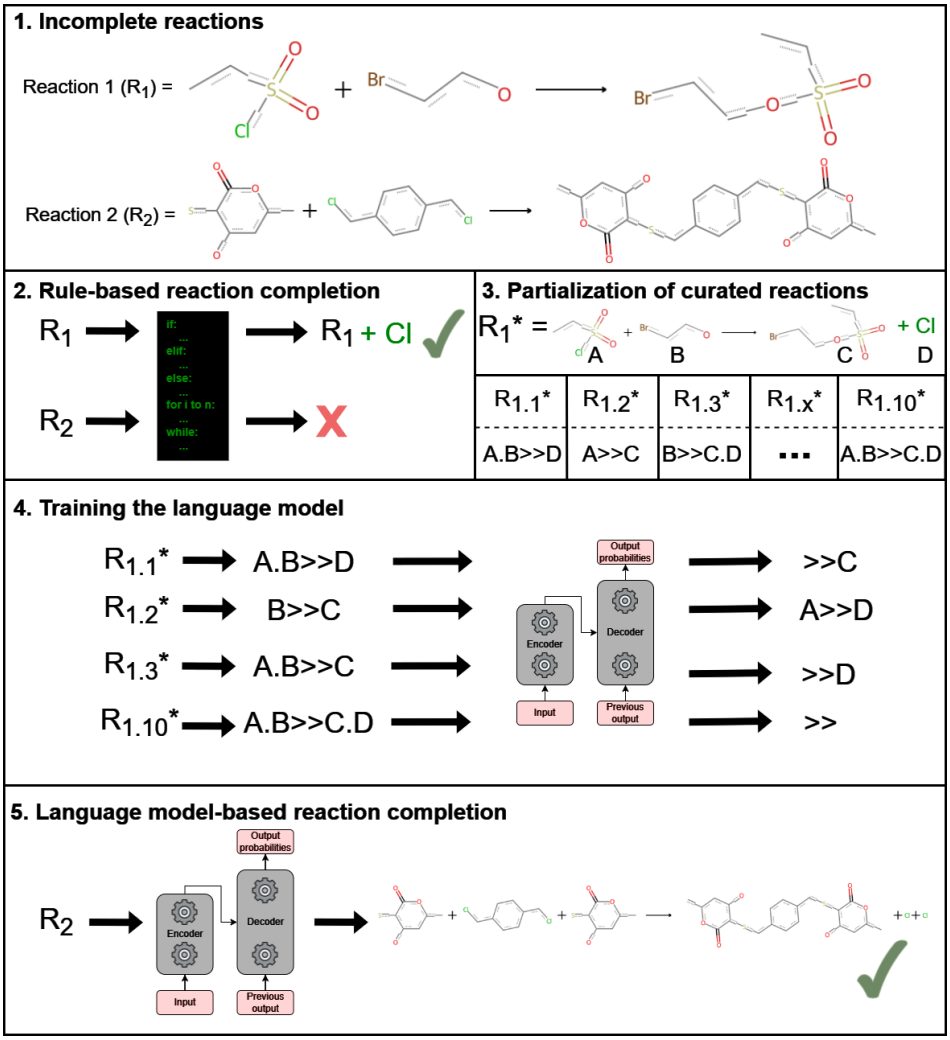

Figure 2: Overview of the methods. **Box 1** shows two example reactions which are incomplete. **Box 2** shows the output of our rule-based method. Some reactions, like R_1, are solved, while others, like R_2, are not. **Box 3** shows the data preparation step for our model-based method. Reactions which are successfully curated using the rule-based method are taken and modified in up to 10 different ways. Molecules are randomly taken out of the complete reaction to create partial reactions, of which the missing set of molecules is known. **Box 4** shows how partialized reactions are used to train our LM. By exposing the transformer to many examples of incomplete reactions, it learns to predict the set of missing molecules. **Box 5** shows how the trained language model is applied to reactions which were previously not completed by the rule-based method.

not curated in this step are usually more complex or are missing crucial components of its reaction. We find that our model-based method performs exceptionally well on our artificially partialized subset of curated reactions, but struggles with remaining uncurated reactions. We present findings which elucidate the strengths and weaknesses of the current approach, such as its effectiveness with or without context molecules, or its accuracy drop as more molecules are missing. Lastly, we measure the correctness of our predictions using a personalized evaluation metric. Our work provides insights and findings which directly affect the quality of research and development on various topics in predictive chemistry and sustainable chemistry.

Methods

Two separate methods are presented in this work that share the same goal of completing incomplete chemical reactions. The first method, the *rule-based method*, uses a set of hard-coded mathematical and chemical rules to analyze incomplete

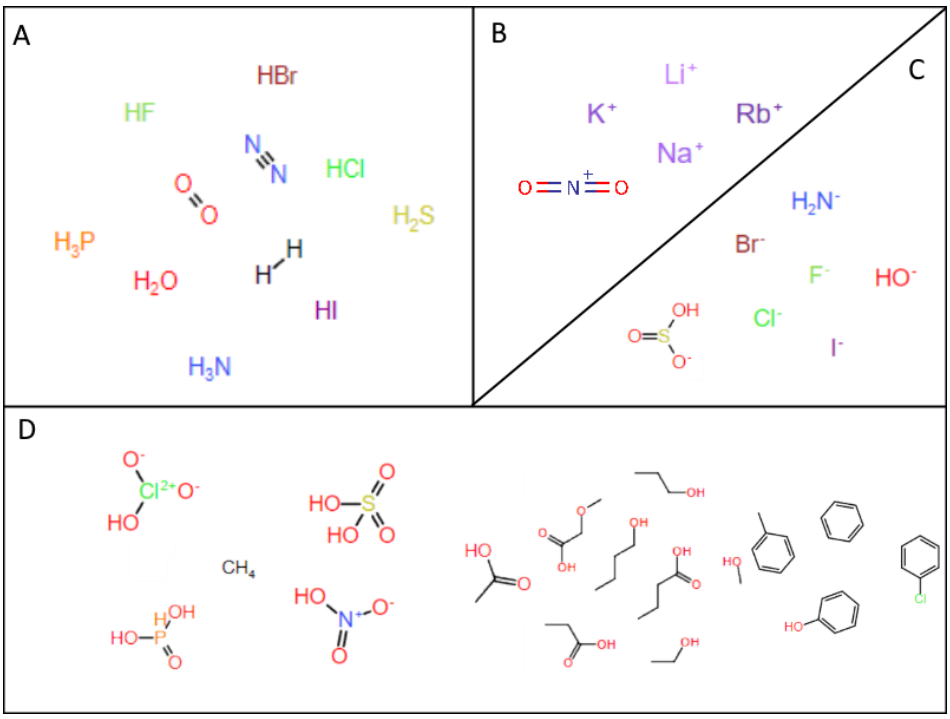

Figure 3: Helper species sets. Set 'A' shows the strict set of neutral helper species molecules. Set 'B' shows the positively charged helper species. Set 'C' shows the negatively charged helper species. Set 'D' shows the lenient set of helper species.

reaction equations and identifies the likely set of missing molecules necessary to output a balanced reaction. The second method uses artificial intelligence, more specifically a transformer-based LM, to learn the language and semantics of chemical reactions. This method exploits the large amounts of data available in the chemical reaction space and attempts to predict the set of missing molecules directly. This method will henceforth be referred to as the *LM-based method*.

The two methods of our hybrid approach are not fully independent. The output of the rule-based method is, after data modification, the data used to train the rule-based method. An overview of our approach is illustrated in Figure 2.

Rule-based Reaction Completion

The rule-based method curates imbalanced reactions using a set of coded rules. Small, commonly found molecules are often identified as missing components in a reaction. Furthermore, missing stoichiometric information is identified too. Usually, a combination of missing molecules and stoichiometry is necessary to balance a reaction.

Definition of balanced reactions. In order to curate a reaction, the definition of a reaction being 'balanced' needs to be properly defined. In this paper, two core rules are considered. First, reactions must have similar amounts of atoms of each type on either side of the reactions. Second, the cumulative charge of molecules on each side of the reaction must be of similar value. When both rules are followed, a reaction is both atomically- and charge-balanced.

Solvent-identification. In order to make a correct atom- or charge-balance of a chemical reaction, solvent molecules (including other non-reacting reagents like catalysts) should not be included in the count. These molecules should either appear on both the left- and right-hand-side (LHS & RHS) of a reaction, or not at all. In this paper, solvents are not taken into account when processed in the rule-based method. Some datasets have reacting and non-reacting molecules mixed on the LHS of the reaction, making identification of a proper atom balance considerably harder. Therefore, only datasets with readily-identified solvents are used.

Helper Species set. The reaction completion algorithm attempts to curate the imbalanced reaction by adding small, frequently-appearing molecules to either side, known in this context as *helper species*. Helper species are a set of molecules that are often missing in the original reaction. Specific combinations of atoms that are missing on one side of the reaction observed significantly more often than others, pointing to a pattern that spe-

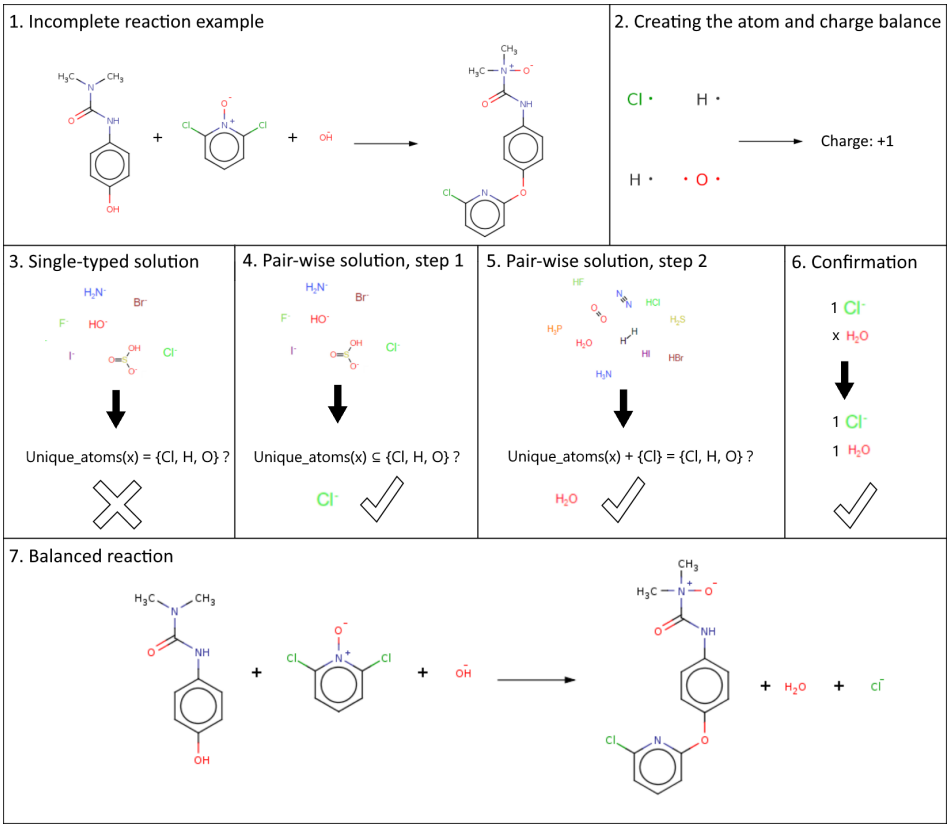

Figure 4: **1.** The original reaction, as found in the dataset. **2.** The identified imbalance: a surplus of one chlorine atom, two hydrogen atoms, and one oxygen atom on the LHS, as well as a charge-imbalance, namely one more positive on the RHS. **3.** Single-helper species balancing attempt. Fails due to no molecule in set C having similar atom types to the atom type-set missing. **4.** Pair-helper species balancing attempt. Looser requirements allows molecules to continue to the next step if its unique atom type-set is a subset of the missing atom type-set, hits first for the chlorine ion. **5.** Using the chlorine ion as a base, pair with molecules in set A to find pairs whose combined unique atom set equals missing atom type-set. Since H and O are missing, H_2O is first to hit. **6.** Balancing algorithm finds coefficients for each helper species. For this charge-based example, the charge-based helper species is 'fixed' at one, due to needing only one to balance the charge imbalance. For charge-balanced examples, a helper species with a unique atom type is 'fixed'. The other molecule's coefficient is easily calculated if it is a correct match. **7.** Curated solution.

cific molecules (which make up the combination of atoms) are missing. The size and content of the helper species set is arbitrary. No consensus exists on what molecules should be manually included to reactions. Previous studies have approached this problem similarly, but with different levels of leniency for the set of helper species [5, 14]. Being strict or lenient on which molecules you allow in the helper species set affects both the rate and accuracy of curation. The wider the range of molecules considered when completing incomplete reactions, the greater the number of reactions will be balanced. However, it also increases the chance of accidentally matching a molecule, increasing the rate of false-positives.

This paper divides the available molecules that comprise the helper species set into four categories, illustrated in Figure 3. Set A is considered the base set of helper species molecules that are frequently observed and have a very low level of ambiguity when encountered in example reactions. Water, gases and hydrogen-acids (such as HF) represent most of this set. Set B and C are helper species meant to balance electronically imbalanced reactions, by including common ions or charged molecules. Lastly, set D is the helper species set referred to as *lenient*. This set contains a greater variety of molecules, ranging from aromatic compounds to alcohols, among others. This set is strongly influenced by the choice of helper species in a different reaction completion paper [14].

Curation using helper species. Determining the correct types and amount of helper molecules is the core of our rule-based algorithm. Previous papers have explored different workflows. Arun et al. attempts to find the correct helper species by iteratively adding candidate molecules which are the closest match to the missing combination [14]. Zhang et al. takes a similar approach which includes a linear solver that identifies solutions requiring stoichiometric changes [5]. This paper approaches helper-species based curation similarly to the aforementioned methods. Helper species are considered most promising additions when their atom types coincide with the imbalanced atom types of the original reaction. When reactions have an atom surplus on the LHS of the equation, molecules are added on the product side. Likewise, reactions with an atom surplus on the RHS are curated by including new molecules on the reactant side. Some reactions contain a surplus of different atom types on either side, requiring additional helper species on both sides. For a step-by-step overview, see Figure 4. In this example, multiple helper species are required to curate the reaction using the rule-based method. Note that this specific example has two valid solutions, namely adding byproducts H_2O and Cl^-, as well as adding byproducts HCl and OH^-, showcasing common occurrences of ambiguous solutions.

Single helper species are first added to the deficit side. For 'strict' curation, molecules contained in set A from Figure 3 are considered. If the reaction is charge-imbalanced, molecules from set B or C are considered. Solutions are found when the addition of one or more of that helper species balances the reaction. When unsuccessful, the algorithm explores solutions requiring two different kinds of helper species. Pairs between molecules of set A are considered, as well as A+B and A+C in case of charge-imbalance. For curation on the 'lenient' helper set, pairs between two molecules of set D are not considered. This decision was based on three reasons. First, the algorithmic complexity encountered when also considering variance in stoichiometry for the helper species increases enormously. Second, a previous paper noted that the wider availability of helper species only marginally improves the curation rates [5]. Third, this increases the rate of ambiguous outputs, which arises from employing a too wide range of possible outcomes, as discussed earlier.

Erroneous reactants. In some rare cases, molecules are erroneously labeled as reactants in the original dataset. This is likely an artifact of the imperfect data-mining software used to generate the large patent-based datasets. To circumvent this problem, the atom balance of an uncurated reaction is compared to the atom balance of each individual reactant. If a perfect match is found, i.e., the types and numbers of atoms missing on the RHS of the equation are exactly equal to the atom type and count of one of the reactants, this reaction likely accidentally labeled a non-reacting molecule as a reactant. The reaction is curated by moving these reactants to the solvent set.

Model-based Reaction Completion

The language model presented in this paper is based on the Molecular Transformer [6], which in turn is based on the original transformer model [16]. As discussed previously, this deep learning model has shown to excel in the field of natural language processing (NLP) and translation tasks. Furthermore, the transformer model has shown to be remarkably successful in the chemical domain. Molecules are presented as string-formatted 'words' using SMILES (simplified molecular input line entry system) notation [18], and reactions as a sequence of words, akin to a sentence. The Molecular Transformer utilizes this strength to achieve incredible results in the field of reaction outcome prediction [6].

Pre-trained and fine-tuned. The Molecular Transformer is an LM fully trained on the chemical 'language' for its forward reaction prediction task [6]. Its non-proprietary accessibility is an opportune privilege for our research, as it skips the need to train a transformer model from scratch. Fine-tuning of the model is necessary in order to shift its problem domain to that of reaction completion. The version of the Molecular Transformer used was one that was trained on augmented data from the USPTO_STEREO set with reactants and solvents separated from each other, averaged over the last 20 checkpoints, named *STEREO_separated_augm_model_average_20.pt*[4]. Fine-tuning was achieved by training the pre-trained model on our partialized reaction data, until no more improvement was observed on the validation set.

Tokenization scheme. The LM receives a set of tokens as input, and produces a set of tokens as output. The type of tokenization scheme employed differs per study. The original Molecular Transformer tokenizes every character of a reaction SMILES string individually, including punctuation characters such as '.' and '>', representing a separation between two molecules and

[4] https://ibm.ent.box.com/v/MolecularTransformerModels

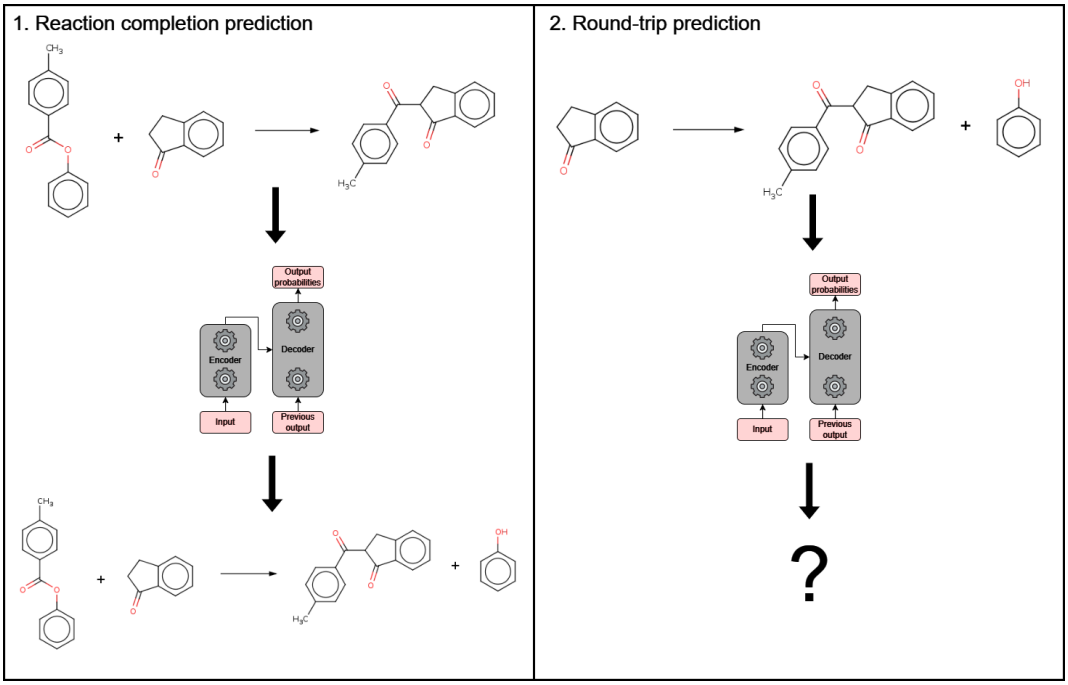

Figure 5: An example of round-trip prediction evaluating the accuracy of our model's prediction. First, the language model predicts the missing molecule for an incomplete reaction. In this example, our model predicts the single aromatic compound as a missing byproduct. This balanced reaction is then deliberately made incomplete again by removing a different set of molecules from the predicted reaction. If the second prediction returns the same reaction as the original prediction, we consider our prediction round-trip accurate.

the separation between reactants/solvents or solvents/products respectively [6]. A more recent reaction completion study experiments with byte-pair encoding (BPE) [5]. BPE uses an algorithm to identify the most commonly occurring substrings, and groups characters together when they are frequently appearing. For example, the sequence 'ccc' describes three carbon atoms linked to each other in an aromatic ring, which are then condensed into a single token. Considering the NLP background of the transformer, the context of commonly occurring atoms can be telling for a molecule's structure. It has been argued that this type of tokenization also helps the model understand molecular disconnection strategies better [5]. However, thus far no study has been conducted that analyzes the effectiveness of different tokenization schemes on the performance of sequence-to-sequence models. In this study, we adopted the single-character tokenization scheme of the Molecular Transformer.

Data format. In order to shift the fine-tuned Molecular Transformer's task from reaction outcome prediction towards missing molecules prediction, the LM needs to train on a different data format. Reactions that were successfully curated by the rule-based method and naturally balanced reactions were stored in a separate curated data file. These reactions were then 'partialized', meaning a number of molecules were deliberately removed from the reaction and placed in a target file. The number of molecules removed - the *degree of partialization* - was restrained in such a way that partial reactions could not have fewer than half the atoms found in the original reaction, as that would be an unfair prediction task. Each reaction was split into up to 10 unique partial reactions. Less than 10 partial reactions were generated in the case of reactions with very few molecules, and thus very few possible combinations of missing molecules. This includes the partial reaction which is equal to the original reaction, in order to teach the model that when faced with already-balanced reactions, no change needs to be made. See box 3 in Figure 2 for an illustration of this process, and box 4 to see how these partial reactions are used to train our model.

The partialized dataset was split into train/test/validation sets on a 90/5/5 split. Due to the multiplication of data during partialization, the size of the dataset was large enough to warrant a smaller validation and test set while maintaining a diverse enough set of reactions in either set. The large amount of data also let us

partialize reactions belonging to the test set only once. On some occasions, two different reactions can produce a similar partialized reaction. When this happens, both correct 'answers' are recorded for each partial reaction using a tab-separated format. During testing, outputting either one of those answers is considered correct.

Evaluation Metrics

It is difficult to verify the correctness of proposed, balanced reactions. As seen in Figure 4, two solutions are viable and indistinguishable in likeliness unless judged by an expert in organic chemistry. In the ideal scenario, every proposed reaction is experimentally verified. However, that is both unfeasible and contrary to the motivation of this study, namely to automate the reaction completion task. Previous reaction-completion studies mention the issue of correctness, but do not present viable methods beyond an atom- and charge-balance check [5, 14, 17].

We propose utilizing a metric previously introduced in a different field of predictive chemistry, the round-tip accuracy [19]. Retrosynthesis, the prediction task of suggesting possible precursor molecules for a product, suffers from a similar lack of ground-truth for data-driven models. In order to increase the confidence that a suggestion is correct, the output precursors are used as input for a separate forward prediction model. If the result of this 'reverse' prediction equals the original product, the prediction is considered round-trip accurate. In order to translate this metric to our problem, the set of missing output molecules cannot be used as the sole input for a second prediction, as they are almost always too few to create a meaningful, partial reaction. To solve this problem, a new partial reaction is created which always includes all the molecules from the suggested output, as well as some other molecules. If the output of the second prediction equals the molecules that were left out of the second input, the prediction is considered round-trip accurate. An illustration of this method is shown in Figure 5.

Results and Discussion

Dataset

The choice of dataset for our research was influenced by a few factors. It needed to be freely accessible, it needed to be large enough to properly train a language model on, and the quality of reactions should not be too low. A large corpus of reaction data was made publicly available using data-mining in 2012, known as the $USPTO$ dataset [2]. This database is freely accessible and large (>3M reactions), but suffers from many bogus or erroneous reactions. Over the years, other studies filtered this database, causing various subsets to be available. The specific subset we used, named $USPTO\_STEREO$, is a filtered version of the originally noisy dataset, but underwent less filtering than other datasets and kept stereochemical information, totalling 1002970 reactions [6]. The $USPTO\_STEREO$ dataset was analyzed on how *complete* the dataset was, which can be seen in Figure 6. Only less than 40k ($< 4\%$) reactions were naturally balanced, while the majority of reactions appeared to have atoms missing on the RHS of the equation (i.e, missing byproducts). The three largest red bars correspond to imbalanced reactions where the byproducts HCl, H_2O and HBr are missing. Early analysis on reaction datasets influenced the choice of most promising helper species.

Rule-based curation

The rule-based curation algorithm balanced a total of 557379 reactions (55.6%) when using the lenient helper set (see Figure 3) and 495197 reactions (49.4%) when using the strict helper set. It is clear that using a wider range of possible helper species molecules curates a greater number of reactions. The vast majority of curated reactions were curated using the helper species module, while around 3.5% of the original dataset was pre-balanced, and a marginal number of reactions were balanced by identifying erroneous reactants ($< 0.5\%$).

Previous papers have reported a variety of curation rates. Arun et al. show a remarkably high rate of curation, but this is placed in the context of a long pipeline with previous operations already filtering out a large number of reactions [14]. Despite that, the number of reactions drop from 55% to 53% of the original size after their 6^{th} step - the step which balances incomplete reactions. The many steps occurring previously are likely reason for the high curation rates, though this warrants further investigation. Zhang et al.'s paper on rule-based curation report a rate of 17.6% [5]. This is in stark contrast to our results, as well as other paper's results. We verified these curation levels by testing their curation algorithm on a similar set of reactions as ours, and found that their algorithm curates 564 of the 1000 example reactions, while the rule-based algorithm presented in this paper curated 590 reactions. We suspect the stark difference might be related to the different way rule-based curation is interpreted by either study. Our

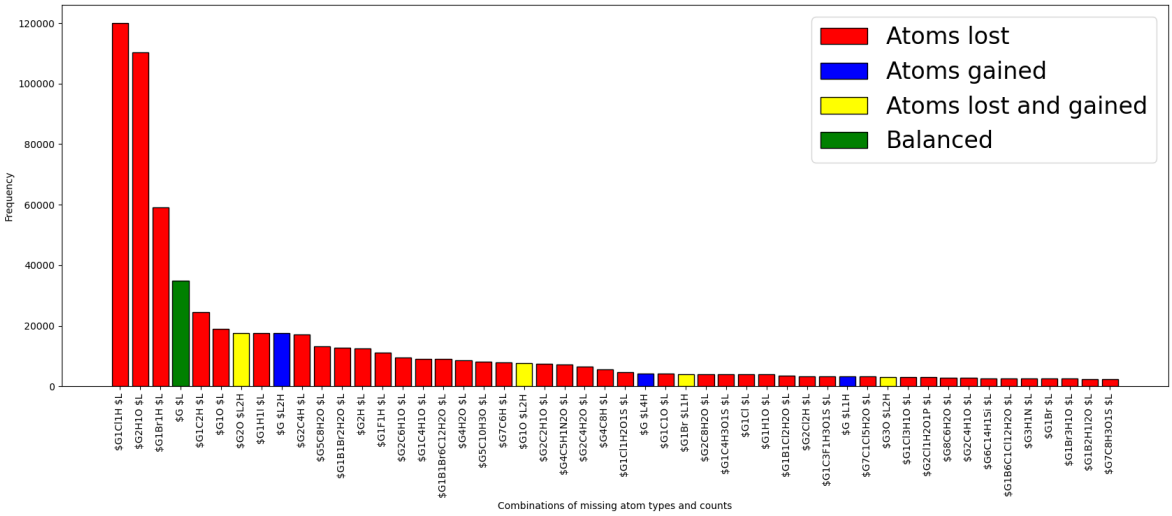

Figure 6: The 50 most common combinations of missing atoms in chemical reactions for the *USPTO_STEREO* dataset. Red bars illustrate combinations where atoms are missing on the RHS, blue bars for those missing atoms on the LHS, yellow bars for a combination of both, and the green bar illustrates the frequency of complete reactions. X-axis labels can be read as always starting with a $\$G$, proceeded by atom types and amounts for atoms missing on the RHS, then a space and a $\$L$ for atom types and amounts missing on the LHS.

hybrid approach sees rule-based curation happen independently, while their rule-based curation is interwoven with its model-based curation.

Model-based curation

When limiting the scope of the reaction incompleteness problem to partial reactions with exactly one molecule missing, our LM achieved a top-1 accuracy of 96.3%. We trained a separate model on exactly the same partial reaction data, but including context molecules like solvents, and found that its accuracy dropped slightly to 96.0%. The model trained on data that includes solvent information contains additional chemical context, while the model trained without this information is of shorter input length. The difference in performance is marginal, but due to the reduced complexity of input data, all further experiments were done using the solvent-excluded model.

Degree of partialization. Figure 7 illustrates the high rates of success that our language model returns on reactions previously curated by the rule-based method. Accuracy rates range between 88% and 95% when considering the top 5 results. Logically, as we increase the beam size of our model, it can explore multiple paths down its decision tree and pick the output sequence with the highest probability. We also analyzed the frequency of how often predictions are valid SMILES strings. That is, how often does the language model actually output something in 'coherent' chemical language. The logarithmic scale of the top right bar graph shows that there is quite a discrepancy between the first and the second-best consideration. We suspect that this might have to do with the model being very sure about a specific answer (top-1), so much so that its second answer is a near-copy of it, except for minor token deviation. In the SMILES language, a minor token deviation can easily invalidate the molecule. Different degrees of partialization are experimented on in the heatmap in the bottom left of the figure. A gradual decrease in accuracy can be seen as we move towards the right side of the heatmap. This translates to harder tasks for our language model, since the model needs to predict a longer sequence, i.e., a more molecules. Interestingly, an unexpected blue square is observed in the bottom-left. We suspect that this group of partial reactions is particularly difficult to decipher due to these reactions being stereoisomerization reactions: one molecule changes its three-dimensional spatial arrangement without losing or gaining atoms. This specific combination occurs 285 times, with an average accuracy of 55.4%, much lower than average. The heat map in the bottom right of the figure helps illustrate how common certain combinations of partial reactions appear. The majority of chemical reactions are around two to four molecules long, with one to two molecules missing.

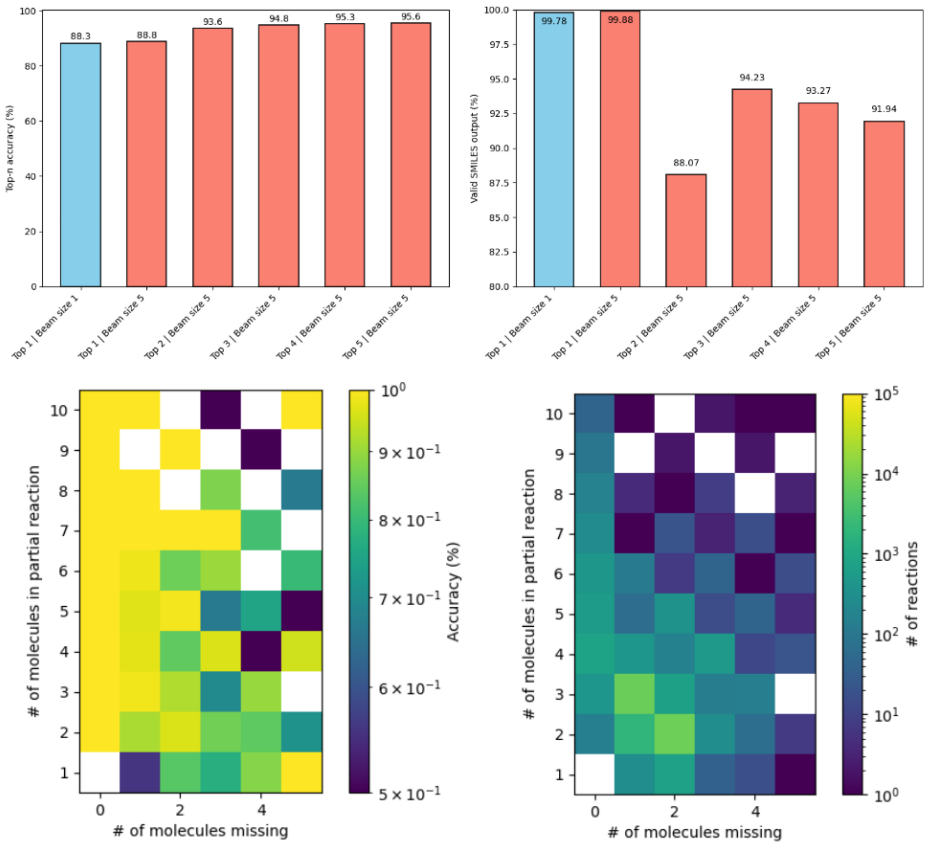

Figure 7: **Top left.** Bar graph showing the accuracy of the language model on partial reactions from the curated subset. The blue bar shows the top-1 accuracy of the model when limiting the beam size to 1. The red bars show the cumulative top-n accuracy as the first n options are considered on a beam size of 5. **Top right.** Bar graph showing the rate of valid SMILES strings, per n^{th} prediction. **Bottom left.** Heatmap showing prediction accuracy on the test set for partial reactions for different combinations of degree of partialization. Notably, the model correctly identifies complete reactions due to the consistent yellow hue on the left column. **Bottom right.** Heatmap showing how frequent specific combinations of degree of partialization are. Note the logarithmic scale. A majority of partial reactions belong to the 2-to-3 group on either axis.

Model type	'Short' accuracy	'Medium' accuracy	'Long' accuracy
ChemBalancer	99.9%	78.3%	16.4%
Our LM-based method	99.9%	91.8%	82.8%

Table 1: Table showing different accuracy rates of different LM's across different lengths of outputs. Both models see accuracy degrade as the length of the solution increases, though our LM-based method degrades significantly less fast.

Compared to the previous paper which fine-tuned the Molecular Transformer, which saw an accuracy of 30.4%, there is a large difference. A couple of factors could be attributed to this [17]. The largest difference is that the other paper considers solvent predictions part of the problem. We decided not to include solvents, as they are not an active part of the reaction.

Length-based performance. A previous study categorized the performance of their language model based on the number of output tokens required in the solution [5]. Zhang et al.'s language model uses a slightly different tokenization scheme, as discussed previously, so in order to make a fair comparison, this needs to be taken into account. In their paper, 'short' solutions are those of token length 1, 'medium' from 1 up until 10, and 'long' greater than 10. As their tokenization scheme sometimes groups up to three or four characters together, the average length of the token is assumed to be 2 characters long. We attempted to find an exact, 'average' length of their tokens, but

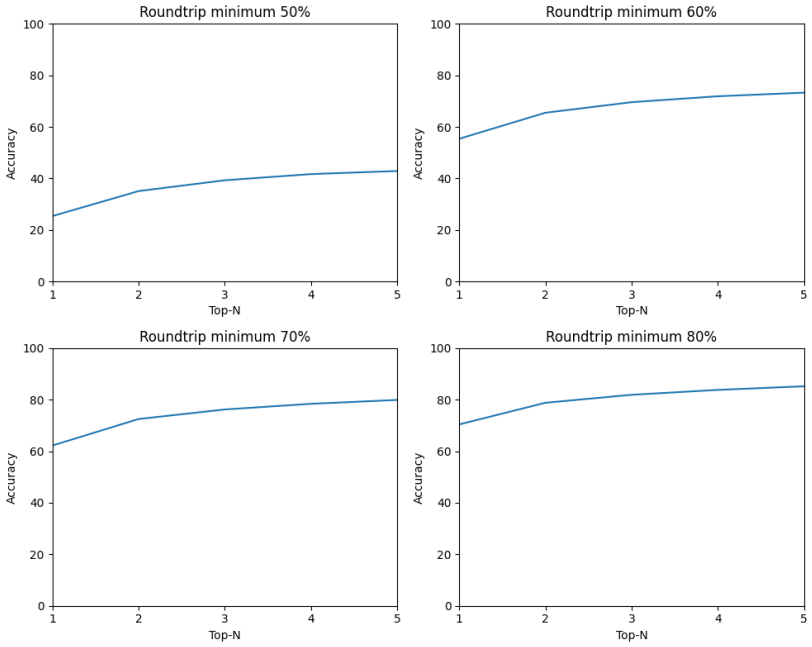

Figure 8: Four different cut-off points for round-trip accuracy. A round-trip minimum of 50% means at least 50% of the atoms in the complete reaction need to appear in the modified, partial reaction. Higher levels of round-trip minimum logically return higher levels of accuracy. Similarly, a higher top-n also directly correlates to a higher round-trip accuracy.

found that the tokenization file was private. Furthermore, because our output scheme is always in the format of '> >' (to separate reactants from products), the predictions belonging to 'short' solutions are at most 7 characters long. This takes the three characters of the mentioned default into account, as well as two more character with two spaces to separate the tokens. 10 token are translated to 20 characters in our case, with an additional 20 space characters. As a result, 'medium' predictions are up until 43 in length, while 'long' predictions are 44 and longer. The results of this analysis are shown in Table 1. The reasoning behind our model's stronger performance on longer outputs likely lies in the very different architecture of the LM used in the other study. Their LM severely limited its flexibility with longer outputs. Meanwhile, our transformer-based architecture degrades much less when the length of the output increases. A similar trend was found when measuring its accuracy against predictions of different total molecular weight. The greater the weight of missing molecules, the further the accuracy degrades.

Application on uncurated reactions

Applying the language model on the set of uncurated reactions shows lesser results. Out of the total of 507773 reactions unsuccessfully curated when using the strict set of helper species, only 27238 partial, unknown-ground-truth-reactions (5.4%) produced an output both atom- and charge-balanced. This is quite a stark contrast to the results obtained in previous sections. We suspect that the difficulty of completing the remaining reactions is much higher. Many of these reactions are missing important components to reasonably predict the missing molecules. At the same time, there is room for improvement. Our LM-based method has been trained on a very specific type of incomplete reaction: reactions which are curated by our rule-based approach. Our rule-based approach is very effective at fixing synthesis reactions, but has little success for decomposition reactions or substitution reactions. Decomposition reactions are reactions where a larger molecule reacts to split into multiple smaller molecules. Substitution reactions see molecules react to swap functional groups with each other. If our LM-based method was more exposed to different types of reactions, this might cause it to generalize better to different types of reactions. Nevertheless, we do note that an additional 5.4% of the remaining subset of incomplete chemical reactions are potentially curated.

Measuring correctness

A modified version of the round-trip accuracy metric was used in an attempt to assess the correctness of the model's predictions. Since the level of partialization in this situation is arbitrary, we experimented the round-trip accuracy across four different values: 50%, 60%, 70% and 80%. This can be seen in Figure 8. As we expected, the more lenient we are with round-trip predictions (i.e., a higher round-trip minimum, so there are less missing molecules), the higher our model predicts the exact same reaction again, using a different partial reaction equation. Similarly, if we allow the model to make multiple predictions for the same partial reaction, we are more likely to find the exact same complete reaction. The largest increase in round-trip accuracy was seen when increasing the value from 50% to 60%, indicating that the former percentage might be too harsh. We consider our model relatively reliable, as we observe round-trip accuracy levels of around 80%. This indicates that the model is more sure about a very specific complete reaction being the one to go to.

While this correctness metric gives us an insight in how reliable our language model is, it is still not a golden solution. Balanced reactions do not necessarily equate with correct reactions. Currently, the only way to truly measure the correctness of a chemical reaction, is to have a chemist look at a subset of predictions. Unfortunately for this research, this is out of our scope.

Discussion & outlook

In this work, we present a hybrid approach to completing incomplete reactions. Inspired by previous developments in the field of predictive chemistry, we made use of the promising results found when using language models to solve problems in the chemical domain. By using the Molecular Transformer, a deep-learning transformer-based language model, we developed a model-based method which is capable of understanding the language of chemistry, and identify what kind, and how many of that molecule is missing in a reaction. Simultaneously, we used a rule-based method to curate the bulk of incomplete reactions by applying logical rules based on atom and charge balances, as well as some limited chemical knowledge. By using the output of the rule-based method to feed and train our model-based method, we present a hybrid approach that managed to curate over half of a commonly used chemical reaction dataset.

We found that our language model performs exceptionally well on our test set, with rates above 95% accuracy. Further analysis showed that this strong performance persists even with more difficult prediction tasks. However, we also found that this performance is in stark contrast to its performance on the uncurated reaction subset, which was only 5.4%. The rule-based method initially failed to curate these reactions, and subsequently did not include these reactions in the training of the model. An experiment on how well our model performs based on different reaction classes could possibly elucidate the reasons for the stark difference in performance. It is likely that certain classes of reactions were difficult to curate using the rule-based method, and were thus underrepresented in the training data.

We went further than other studies in measuring the correctness of our predictions by using a round-trip accuracy metric. using this correctness measure, we found that the model is likely to produce a similar complete reaction output when the question is posed in a slightly different manner. This is promising, but also not definitive for its correctness. We acknowledge that the best way to assess correctness of completing incomplete reactions is by manually checking predictions, which is both cumbersome and out of scope for this research.

For future work, we propose exposing the language model to a wider variety of complete chemical reactions. This can be achieved by employing a more thorough rule-based method. Having an expert chemist design and encode important chemical knowledge can help broaden the types of reactions curated, which indirectly improves the model's performance.

In conclusion, this paper presents a hybrid approach to the reaction incompleteness problem. We find that a significant portion of reaction databases can be curated using our method. However, some types of reactions are likely underrepresented in the curation effort, which leaves room for future improvement.

References

[1] Philipp-Maximilian Jacob and Alexei Lapkin. Statistics of the network of organic chemistry. 3 (1):102–118. ISSN 2058-9883. doi: 10.1039/C7RE00129K. URL https://pubs.rsc.org/en/content/articlelanding/2018/re/c7re00129k. Publisher: The Royal Society of Chemistry.

[2] Daniel Mark Lowe. Extraction of chemical structures and reactions from the literature.

[3] Jana M. Weber, Zhen Guo, Chonghuan Zhang, Artur M. Schweidtmann, and Alexei A. Lapkin. Chemical data intelligence for sustainable chemistry. 50(21):12013–12036. ISSN 1460-4744. doi: 10.1039/D1CS00477H. URL https://pubs.rsc.org/en/content/articlelanding/2021/cs/d1cs00477h. Publisher: The Royal Society of Chemistry.

[4] John A. Keith, Valentin Vassilev-Galindo, Bingqing Cheng, Stefan Chmiela, Michael Gastegger, Klaus-Robert Müller, and Alexandre Tkatchenko. Combining machine learning and computational chemistry for predictive insights into chemical systems. 121(16):9816–9872. ISSN 0009-2665. doi: 10.1021/acs.chemrev.1c00107. URL https://doi.org/10.1021/acs.chemrev.1c00107. Publisher: American Chemical Society.

[5] Chonghuan Zhang, Adarsh Arun, and Alexei Lapkin. Completing and balancing database excerpted chemical reactions with a hybrid mechanistic - machine learning approach. URL https://chemrxiv.org/engage/chemrxiv/article-details/64b7feb7ae3d1a7b0dfeba80.

[6] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. 5(9):1572–1583, . ISSN 2374-7943. doi: 10.1021/acscentsci.9b00576. URL https://doi.org/10.1021/acscentsci.9b00576. Publisher: American Chemical Society.

[7] Bowen Liu, Bharath Ramsundar, Prasad Kawthekar, Jade Shi, Joseph Gomes, Quang Luu Nguyen, Stephen Ho, Jack Sloane, Paul Wender, and Vijay Pande. Retrosynthetic reaction prediction using neural sequence-to-sequence models. 3(10):1103–1113. ISSN 2374-7943. doi: 10.1021/acscentsci.7b00303. URL https://doi.org/10.1021/acscentsci.7b00303. Publisher: American Chemical Society.

[8] Philippe Schwaller, Alain C. Vaucher, Teodoro Laino, and Jean-Louis Reymond. Prediction of chemical reaction yields using deep learning. 2(1):015016, . ISSN 2632-2153. doi: 10.1088/2632-2153/abc81d. URL https://dx.doi.org/10.1088/2632-2153/abc81d. Publisher: IOP Publishing.

[9] Hanyu Gao, Thomas J. Struble, Connor W. Coley, Yuran Wang, William H. Green, and Klavs F. Jensen. Using machine learning to predict suitable conditions for organic reactions. 4(11):1465–1476. ISSN 2374-7943. doi: 10.1021/acscentsci.8b00357. URL https://doi.org/10.1021/acscentsci.8b00357. Publisher: American Chemical Society.

[10] Krishna N. Ganesh, Deqing Zhang, Scott J. Miller, Kai Rossen, Paul J. Chirik, Marisa C. Kozlowski, Julie B. Zimmerman, Bryan W. Brooks, Phillip E. Savage, David T. Allen, and Adelina M. Voutchkova-Kostal. Green Chemistry: A Framework for a Sustainable Future. *Organic Process Research & Development*, 25(7):1455–1459, July 2021. ISSN 1083-6160. doi: 10.1021/acs.oprd.1c00216. URL https://doi.org/10.1021/acs.oprd.1c00216. Publisher: American Chemical Society.

[11] Viknesh Andiappan, Andy S. Y. Ko, Veronica W. S. Lau, Lik Yin Ng, Rex T. L. Ng, Nishanth G. Chemmangattuvalappil, and Denny K. S. Ng. Synthesis of sustainable integrated biorefinery via reaction pathway synthesis: Economic, incremental enviromental burden and energy assessment with multiobjective optimization. 61(1):132–146. ISSN 1547-5905. doi: 10.1002/aic.14616. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aic.14616. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aic.14616.

[12] Steven M. Kearnes, Michael R. Maser, Michael Wleklinski, Anton Kast, Abigail G. Doyle, Spencer D. Dreher, Joel M. Hawkins, Klavs F. Jensen, and Connor W. Coley. The open reaction database. 143(45):18820–18826. ISSN 0002-7863. doi: 10.1021/jacs.1c09820. URL https://doi.org/10.1021/jacs.1c09820. Publisher: American Chemical Society.

[13] Timur R. Gimadiev, Arkadii Lin, Valentina A. Afonina, Dinar Batyrshin, Ramil I. Nugmanov, Tagir Akhmetshin, Pavel Sidorov, Natalia Duybankova, Jonas Verhoeven, Joerg Wegner, Hugo Ceulemans, Andrey Gedich, Timur I. Madzhidov, and Alexandre Varnek. Reaction data curation i: Chemical structures and transformations standardization. 40(12):2100119. ISSN 1868-1751. doi: 10.1002/minf.202100119. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.202100119. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.202100119.

[14] Adarsh Arun, Zhen Guo, Simon Sung, and Alexei A. Lapkin. Reaction impurity prediction using a data mining approach**. 3(6):e202200062. ISSN 2628-9725. doi: 10.1002/cmtd.202200062. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cmtd.202200062. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cmtd.202200062.

[15] Vignesh Ram Somnath, Charlotte Bunne, Connor W Coley, Andreas Krause, and Regina Barzilay. Learning graph models for retrosynthesis prediction.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. URL http://arxiv.org/abs/1706.03762.

[17] Alain C. Vaucher, Philippe Schwaller, and Teodoro Laino. Completion of partial reaction equations. URL https://chemrxiv.org/engage/chemrxiv/article-details/60c75240469df41e0ff44b27. ISSN: 1327-3310.

[18] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28 (1):31–36, February 1988. ISSN 0095-2338. doi: 10.1021/ci00057a005. URL https://pubs.acs.org/doi/abs/10.1021/ci00057a005. Publisher: American Chemical Society.

[19] Philippe Schwaller, R Petraglia, VH Nair, and Teodoro Laino. Evaluation metrics for single-step retrosynthetic models. In *Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019)*, 2019.